MULTI-PHASE SAMPLING IN CENSUSES AND SURVEYS

by

CHARLES ALBERT BENDER

B.S., Kansas State University 1965

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Approved by:

A. M. Feyerherm
Major Professor

TABLE OF CONTENTS

INTRODUCTION

Multi-phase sampling is a technique employed by researchers to obtain estimates of parameters using information from previous samples of the same population. The most familiar agency to employ these techniques is the United States Government. There, samples are drawn year after year from the same group of elements, the people of the United States. Multi-phase sampling finds perhaps its widest application in censuses used to estimate totals, average values, and the change in both.

This report is concerned with estimation of parameters in a population using data from field surveys. Problems examined include estimation of the mean of a character, when its procurement is highly expensive, or very difficult and determination of the change in a mean over a time interval. The first problem involves a static population, the second involves a changing population. The first is resolved using double sampling techniques, in which estimates of parameters for one character are made more precise using information from a second, correlated character. The second is resolved using successive sampling techniques, in which information is used from more than the two occasions to be used to estimate the change in the parameter. Several methods for the latter are developed using estimates from totally new samples, fixed samples, and slightly altered samples.

In both areas of multi-phase sampling, methods are developed to obtain the highest precision possible under certain constraints. Methods are also developed to ease sampling and calculation difficulties while maintaining reasonably high precision.

DOUBLE SAMPLING

Sample Allocation: A Simple Case

Suppose in collection of data expensive techniques must be employed to obtain sample values which are used to estimate a certain parameter. The high cost involved precludes a large sample. In turn, a small sample precludes extremely precise estimates. Suppose the character of interest has a known relationship to another character more cheaply evaluated. With methods developed in the following work more precise estimates may be obtained using the relationship than would have been obtained had the appropriation been expended in estimating the parameter of the character of interest alone. Throughout this report one of the primary concerns will be to obtain highest precision with a certain cost.

In the case of double sampling, the most important problem is the proper allocation of sample elements to the two samples. The first sample is involved in estimating the parameter of the character Y of secondary interest. The population is then stratified according to Y. From the strata thus established a subsample is drawn which is used to find an estimate of $\bar{X}$ the mean of the character of primary interest X.

The population is of size N with N_i elements in each of the i, $i = 1, 2, \ldots, s$, strata. The proportion P_i of the population in stratum i is

$$P_i = N_i/N .$$

The first step in double sampling consists of randomly drawing n elements from the N elements of the population. A proportion of this sample p_i, falls into each of the i strata where

$$p_i = n_i/n ,$$

with n_i elements in each such stratum. The second step in the process consists of randomly choosing m_i elements from the n_i elements in each stratum. The second sample is of size m where

$$m = m_1 + m_2 + \ldots + m_s .$$

Let C be the total expenditure of the survey alloted for sampling and A and B be the cost of obtaining a sample value for character X and Y respectively, then

$$Am + Bn = C .$$

In each stratum a sample mean for character X will be obtained, denoted as $\bar{x}_i$. The overall sample mean $\bar{x}$ for X will be

$$\bar{x} = \sum_i p_i \bar{x}_i .$$

An individual sample value j for character X stratum i will be denoted by x_{ij}. An unbiased estimate of $\bar{x}_i$ is

$$\frac{1}{m_i} \sum_j x_{ij} .$$

To find the optimum allocation for n and m a linear function of the x_{ij} and p_i must be found which is an unbiased estimates of $\bar{X}$ with variance smaller than any other like function. There are infinitely many functions whose expected value is $\bar{X}$, of the form

$$F = \sum_{i=1}^{s} \sum_{j=1}^{s} \sum_{k=1}^{m_i} c_{ijk} p_i x_{ij}$$

Where the variance of F is given by:

$$V(F) = \sum_{i=1}^{s} \frac{s_i}{m_i} \left(p_i^2 + p_i q_i n^{-1} \right)^2 + \frac{1}{n} \sum_{i=1}^{s} p_i (\bar{x}_i - \bar{x})^2$$

the requirement of minimum variance necessitates that the initial sample size n be estimated by

$$n' = Cb/(a\sqrt{AB} + bB),$$

and the second sample size m be estimated by

$$m' = Ca/(aA+b\sqrt{AB}) .$$

Where,

$$a = \sum_{i=1}^{s} p_i s_i$$

$$b^2 = \sum_{i=1}^{s} p_i (\bar{x}_i - \bar{x})^2 .$$

Denote the variation of x within stratum i as s_i, where, as before, C is the total appropriation for the survey while A and B are the amounts to be allocated to collecting data on X and Y respectively. Since the values for n' and m' above are estimates and n and m are integers, these estimates are improved by using a' as defined by

$$a' = \sum_{i=1}^{s} s_i \sqrt{p_i^2 + p_i q_i n^{-1}}$$

instead of a in n' and m'. Then substitute these values back into V(F) and check whether the variance is smaller than with the original values.

Sample Allocation: A General Case

Jambunathan (1960) has considered the more special case where the cost of taking an observation varies from stratum to stratum. Denote the cost of sampling from stratum i as A_i, then Neyman's optimum allocations become

$$m_i' = C\ p_i\ s_i/(a_i+Bb)\ \sqrt{A_k}$$

and

$$n' = Cb/(a_i+Bb)\ \sqrt{B}$$

in the general case, where

$$a_i = p_i\ s_i\ A_i$$

and

$$b^2 = \sum p_i\ (\bar{x}_i-\bar{x})^2$$

Under this general case of optimum allocation, the minimum variance V_m is given approximately by

$$V_m = \frac{(a_i+b\ \sqrt{B})^2}{C}$$

However, it is possible to have m_i greater than n_i the expected size of the i-th stratum. In which case no subsampling occurs in that stratum in the second sample. This occurs when

$$\frac{s_i}{A_i} > \frac{b^2}{B}$$

In this event optimum allocation to the remaining strata must be reworked using new population and sample sizes deleting all such strata sizes.

Efficiency of Double Versus Simple Random Sampling

No proof has been presented nor any estimate been made as to the advantages of using double sampling over simple random sampling. To compare

the two their relative efficiency must be calculated as though they were applied with the same expenditure. To begin, redefine A_i, the cost of sampling from stratum i, as the product of B the cost of measuring Y, and the cost of measuring X in the i-th stratum

$$A_i = BD_i^2$$

where D_i^2 is the unit cost of measuring x_{ij}. The minimum variance, V_m, then becomes

$$V_m = B \frac{\sum p_i s_i D_i + \sum p_i(\bar{x}_i - \bar{x})^2}{C}$$

$$= B \frac{(\bar{s}_D + r s)^2}{C}$$

where $\bar{s}_D$ is a weighted mean of strata standard errors; that is,

$$\bar{s}_D = \sum p_i D_i s_i$$

and r is the correlation coefficient of X and Y. How efficient this is in relation to the variance under simple random sampling, remains to be shown.

Employing simple random sampling to estimates $\bar{X}$ means that the cost A of each sample element x_{ij} determines a sample of size (C/A). Therefore, under random sampling,

$$V_{ran} = s^2 A/C = B^2 D^2 s^2/C$$

where D^2 is the ratio of A to B. Denote by E the relative efficiency of double sampling to random sampling, then

$$E = \frac{V_m}{V_{ran}} = \left(\frac{\bar{s}_D}{Ds} + \frac{r}{D}\right)^2 .$$

Double Sampling, with subsequent stratification, will produce gains in precision if $E < 1$, or if $r < D-(\bar{s}_D/s)$. If all the D_i's are the same, i.e., the cost of sampling in all strata is the same, and commonly equal to D, the efficiency becomes

$$D = \frac{\bar{s}}{s} + \frac{r}{D}^2 ,$$

where $\bar{s}$ is the weighted mean of strata standard deviations and s is the overall sample standard deviation. In this case, gains will take place when

$$r < D(1-\frac{\bar{s}}{s}) \text{ or if } D > \frac{r}{1-\frac{\bar{s}}{s}}$$

Thus the efficiency of the design is dependent upon the correlation of the two characters, X and Y. The efficiency is also a function of the relative cost of measuring X as compared to the cost of measuring Y and the ratio of the weighted mean of strata standard deviations and the overall sample standard deviation.

This is as one would have suspected beforehand. If the correlation of the two characters were small, Y would give no information about X and stratification according to Y after the first sample would be ambiguous as to X. Also, if the cost of measuring character Y is approximately the same as that for character X, there would be little point in stratifying and introducing additional complications. This also applies to the standard error ratio. If the two standard errors are not sufficiently diverse, the

cost differential must of necessity be great, perhaps larger than is practically possible.

Ratio Estimates in Double Sampling

In Neyman's (1938) original discussion, only one estimate of the parameter of high cost, $\bar{X}$ was explored. Sukhatme (1962) explored the possibility of more than the one estimator. The three proposed by Sukhatme are all of the ratio type. Recall that from a population of size N, n sample elements were randomly chosen which served to determine the parameter of the less expensive character, $\bar{Y}$. Using these sample estimates to stratify the population as to levels of Y, m_i elements were randomly chosen from the n_i in the i-th stratum, for a total of m elements in the second phase. From these m elements the parameter $\bar{X}$ was estimated.

Let the ratio r_i be defined as

$$r_i = \frac{x_i}{y_i},$$

and the means $\bar{y}_n$, $\bar{y}_m$, $\bar{x}$ and $\bar{r}$ as

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \bar{y}_m = \frac{1}{m}\sum_{i=1}^{m} y_i, \quad \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \bar{r} = \frac{1}{m}\sum_{i=1}^{m} r_i,$$

where $\bar{y}_m$ is the sample mean of the y_{ij} included in the second sample and $\bar{y}_n$ is that of the y_{ij} in the first. Consider the following ratio-type estimates

$$\bar{x}_1 = \frac{\bar{x}}{\bar{y}_m} \cdot \bar{y}_n,$$

$$\bar{x}_2 = \bar{r} \cdot \bar{y}_m,$$

$$\bar{x}_3 = \bar{r}\,\bar{y}_m + \frac{m(n-1)}{n(m-1)}(\bar{x}-\bar{r}\,\bar{y}_m) .$$

The first two are biased, while $\bar{x}_3$ is an unbiased estimate of $\bar{X}$, the population parameter.

Using methods developed by Tukey (1956) and extended by Robson (1957) and assuming that N is infinitely large, Sukhatme (1962) shows that the variances of the three estimates are given by

$$v(\bar{x}_1) = \frac{s_x^2 + \frac{\bar{x}^2}{\bar{y}_m^2} s_y^2 - 2\frac{\bar{x}}{\bar{y}_m} s_{xy}}{m} + \frac{2\frac{\bar{x}}{\bar{y}_m} s_{xy} - \frac{\bar{x}^2}{\bar{y}_m^2} s_y^2}{n},$$

$$v(\bar{x}_2) = \frac{\bar{y}_m^2 s_r^2}{m} + \frac{\bar{r}^2 s_y^2 + 2\bar{r}\bar{y}_m s_{ry}}{n},$$

$$v(\bar{x}_3) = \frac{s_x + \bar{r}^2 s_y^2 - 2\bar{r} s_{xy}}{m} + \frac{\bar{r}^2 s_y^2 + 2\bar{r}\bar{y}_m s_{ry} + 2\bar{r}\,E(y-\bar{y}_m)(r-\bar{r})}{n} .$$

In each of the variance equations the estimate is composed of two terms. The first term represents the error contributed by the m sample elements form the strata as though n, the initial sample size, were exactly N, the

population size. The second term represents the error due to the initial sample which is a simple random sample of the population of size N.

Since $\bar{x}_3$ is an unbiased estimate of $\bar{X}$, and $\bar{x}_1$ and $\bar{x}_2$ are biased, one might wish to find the relative efficiency of $\bar{x}_3$ as compared to $\bar{x}_1$ and $\bar{x}_2$. Define

$$b_1 = \frac{E(r-\bar{r})^2 (y-\bar{y}_m)}{E(r-\bar{r})(y-\bar{y}_m)},$$

and

$$b_2 = \frac{E(r-\bar{r})(y-\bar{y}_m)^2}{E(r-\bar{r})(y-\bar{y}_m)}.$$

Sukhatme shows from the difference between variances that $v(\bar{x}_3)$ will be smaller than $v(\bar{x}_2)$ when

$$b_1 < -\frac{1}{2\bar{y}_m} \quad \text{and} \quad b_2 < -\frac{1}{2\bar{r}}.$$

If these conditions are satisfied, $\bar{x}_3$ is superior to $\bar{x}_2$ in that $\bar{x}_3$ is unbiased and its variance is smaller. When these conditions are not satisfied, $\bar{x}_2$ is not, of necessity, preferrable to $\bar{x}_3$ because $\bar{x}_2$ is biased. The amount of the bias must be weighed against the gain in precision to decide which is preferrable.

Comparing the difference in variances, $\bar{x}_3$ is more precise than $\bar{x}_1$ if

$$0 < \bar{x}/\bar{y}_m < \bar{r} \leq b, \text{ and } b_2 < 0.$$

Regression Estimates in Double Sampling

To utilize more fully available information, Cochran (1963) turned to estimates of the regression of $\underline{X}$ on $\underline{Y}$, where X is the more difficult and costly variable to measure. Assume that a random sample of size $\underline{n}$ is chosen from an infinite population and that the relation $\underline{R}$ between $\underline{X}$ and $\underline{Y}$ is linear. The model will be

$$x_{ic} = \bar{X} + B(y_i - \bar{Y}) + e_{ic} ,$$

with the subscript $\underline{c}$ used as a reminder that for fixed y_i, the random variate e_{ic} is distributed with mean 0 and variance

$$S_e^2 = S_x^2(1 - R^2) .$$

Measuring only character $\underline{Y}$ in the first sample of size $\underline{n}$ and characters $\underline{X}$ and $\underline{Y}$ in the second sample of size $\underline{m}$, the estimate of $\bar{X}$ becomes

$$\bar{x}' = \bar{x} + b(\bar{y}_1 - \bar{y}_2) .$$

Here, $\bar{x}$ is the estimate of $\bar{X}$ from the second sample only, $\bar{y}_1$ and $\bar{y}_2$ are the mean values of character $\underline{Y}$ from the first and second samples respectively, and $\underline{b}$ is the least squares regression coefficient of x_{ic} on y_i. From the model, the error mean square $MSE(\bar{x}')$ becomes

$$MSE(\bar{x}') = \frac{S_x^2(1-R^2)}{M} + \frac{S_x^2(1-R^2)(\bar{y}_1-\bar{y}_2)^2}{\sum(\bar{y}_{2i} - \bar{y}_2)^2} + B(\bar{y}_1 - \bar{y})^2 .$$

If the samples are independent and assumed normally distributed the approximate variance of $\bar{x}$ involves two terms; the first consists of the

variance in the second sample and the second consists of the variance in the first sample with respect to X. The variance of $\bar{x}'$ is approximated by

$$V(\bar{x}') \doteq \frac{S_x^2(1-R^2)}{m} + \frac{R^2 S_x^2}{n} ,$$

The problem of finding the optimum allocation of n and m is the same as in double sampling for stratification. The cost of including an element from the original sample to measure character Y, is c_n and that for including an element in the second sample, used to measure character X and the relation between X and Y, is C_m and C is the total expenditure. Thus,

$$C = nc_n + mc_m .$$

For fixed cost the minimum variance is given by:

$$V_{min} = S_x^2 \left(\sqrt{(1-R^2)C_m} + R\sqrt{C_n} \right)^2 / C ,$$

assuming R positive.

If all the allocation is used to measure X alone, the sample size will be $(m' = C/C_m)$ and the variance V_s of the estimate of $\bar{X}$ under simple random sampling will be

$$V_S = S_x^2/m' = C_m S_x^2/C$$

V_{min} is superior to V_S if

$$C_m/C_n > R^2 \,/\, (1 - \sqrt{1 - R^2})^2$$

or

$$R^2 > 4C_mC_n/(C_m + C_n)^2$$

If the relation between X and Y is known, the first of these equations reveals the critical ratio of the costs, below which no gain in precision is made by double sampling with regression. If one knows the respective costs associated with measuring X and Y, the second equation reveals the critical value of the square of the relation between X and Y, below which no gain in precision is made with double sampling with regression.

Recall that

$$V(\bar{x}') \doteq \frac{S_x^2(1-R^2)}{m} + \frac{R^2S_x^2}{n}$$

if the terms in (1/m) are negligible. With a model of linear regression

$$s_{x\cdot y}^2 = \frac{(n-1)\, s_x^2 - b^2(n-1)\, s^2y_2}{n-2}$$

is an unbiased estimate of $S_x^2(1-R^2)$. Since, s_x^2 is an unbiased estimate of S_x^2 then $(s_x^2 - s_{x\cdot y}^2)$ is an unbiased estimate of $R^2S_x^2$.

An unbiased estimate of $V(\bar{x}')$ is

$$v(\bar{x}') = \frac{s_{x\cdot y}^2}{m} + \frac{s_x^2 - s_{x\cdot y}^2}{n} \quad .$$

Double Sampling Generalized

The case of two phase sampling with the first phase used to stratify the population with respect to the first and inexpensive variable Y, and the second sample used to estimate a parameter of a more expensive variable X, can be expanded further. The next logical progression is to the case of three variables, X, Y, Z. Suppose that character Z is the most expensive and/or difficult of the three to measure followed by X, then Y. The sampling is carried out upon a finite population of size M, each element of which can be expressed with respect to character X_j, Y_i, Z_h as M_{ijh}. The initial sample consists of N elements which serves to stratify the N elements with respect to Y into s strata of size $N_i, i = 1, 2, \ldots, S$. From these s strata n_i elements are chosen randomly from the i-th strata composing the second sample of size n. The n_i elements of sample two, in the i-th stratum, are stratified with respect to X. Within the j-th substratum of stratum i, m_{ij}, $j = 1, 2, \ldots, n_i$, elements are randomly chosen from the n_{ij} elements to measure character Z. The fixed sample sizes N, n, and m are determined by

$$m = m.. \ , \quad n_i = n_i. \ , \quad n = n. \ , \quad N = N. \ .$$

The dot notation indicates summation over all values of the subscript replaced. Appropriate strata proportions are p_i an estimate of the proportion of the population in stratum i after the first sample and p_{ij} an estimate of the proportion in the j-th substratum of stratum i after the second sample. Thus,

$$p_i = n_i/N_i \doteq n/N \ , \quad p_{ij} = m_{ij}/m \ .$$

In general equality does not exist in the above sampling rate estimates since the rates are integers. Define the statistic z_h as an unbiased estimate of the mean of the h-th attribute of character Z, i.e.,

$$N_i n_{ij} m_{ijh} / N n_i m_{ij}$$

$$z_h = \sum_i \sum_j N_i \, n_{ij} \, m_{ijh} / N n \, m_{ij} \ ,$$

subject to the condition that $n_i = 0$ if and only if $N_i = 0$, and $m_{ij} = 0$ if and only if $n_{ij} = 0$. Define $\bar{z}$, the unbiased estimate of $\bar{Z}$, as the average of the z_h.

Because of the inexactness of the strata proportion estimates p_i and p_{ij}, an exact formula for the variance of $\bar{z}$ cannot be written. If one proceeds under the assumption of equality of the strata proportion estimates and the exact values, the only bias in the estimate of the variance of $\bar{z}$ will be the error due to the assumption. Under this assumption Robson and King (1952) obtained very, complex estimates of the variance of $\bar{z}$ which will not be stated here because of its complexity. If k were the size of a single sample taken to measure $\bar{z}$ and the cost of taking this sample were the same as that for a measure of $\bar{z}$ using three phase sampling, then a comparison of the precision of the two dictates that

$$\frac{1}{k} > 1 - \frac{(N-n)\phi^2}{N} - \frac{(n-m)g^2}{n} \ ,$$

where ϕ^2 is Pearsons mean square contingency and g^2 the multiple coefficient of association, both at least zero, but no greater than 1.

From the three phase sampling for stratification and substratification, one can generalize to multi-phase sampling with four or more variables.

Example

An example of the latter, three-phase, sampling scheme is the Curtis Study. This was a survey conducted to evaluate the effect magazines have on their readers. The three characters to be measured in the order of their increasing complexity of measurement were: (1) characteristics of reader families (interests, possessions, buying habits, etc.), (2) characteristics of the readers, and (3) the effect the magazines have upon the readers.

The sampling for the survey consisted of multi-stage probability samples of about 31,300 households. These were based upon 833 sampling units averaging 37 households per unit. The primary sample served to stratify the population with regard to open country, village, or city. This primary sample was the United States Census of 1950 which yielded the appropriate strata proportions. These strata were substratified, to improve precision by using more homogenous groups, by size of city and geographic area under the city stratum and by livelihood, soil type, and agriculture type under both village and open country strata. Double sampling was employed to measure the family and individual characteristics. Under this scheme, the sample units of 37 households were chosen randomly from each of the substrata derived from the primary sample and from this secondary sample of households, individuals were chosen randomly to be interviewed.

SUCCESSIVE SAMPLING

Explanation

Another area of multi-phase sampling is the sampling of the "same" population on successive occasions. Many tables are published monthly, yearly, or every decade, consisting of data concerning single populations. Every civilized country conducts surveys and censuses about everything from artichoke heart consumption to zipper life expectancy. The United States Bureau of the Census devotes many millions of dollars to compiling figures gleaned from a representative sample of the people of the United States about many subjects. More reliance is put on these figures, due to the wider acceptance of sampling and its advantages. But a census every ten years is of limited use because the population of this country is not static since a new person is added every 10-20 seconds and one deleted every 20-30 seconds.

When one samples a population repeatedly, much useful insight is gained about population parameters. This is gained by finding estimates of the mean and variances and perhaps other parameters, time and again which allows the surveyer to project or extrapolate to some future time. But how often and how much should the sample be changed to maintain reliable estimates of the present population's parameters? People are often reluctant to be interviewed repeatedly concerning the same subject. There is a certain amount of feedback from the survey, and the interview itself, which tends to bias the sample. Hence the sample is not necessarily a cross-section of the population. In opposition is the case where the second, and successive, interviews yield more precise estimates than the first, where little interaction between interviewer and subject is possible.

Apart from double sampling techniques already discussed, which are applicable to this discussion as well, there are three kinds of quantities for which estimates may be desired: (1) the changes in $\bar{Y}$ from occasion to occasion, (2) the average of $\bar{Y}$ over all occasions, and (3) the average value of $\underline{Y}$ for the last survey. If the character of interest changes rapidly with time, case three is most applicable. If the change is slow with respect to time, case two is most applicable, using averages of several up-to-date surveys. When the change might be due to some new stimulus or a change in some attribute effecting the population, case one is more relevant.

If it is possible to change the composition of the sample, while maintaining constant sample size, greatest precision will result for the individual cases when; for case (1), to estimate the change in $\bar{Y}$, retain the same sample; for case (2), to estimate the average over all occasions, draw a new sample; for case (3), to estimate the most current $\bar{Y}$, either retain the same sample for all occasions or replace it entirely with a new sample. In the latter case, partial replacement may be more advantageous than either of those above. Partial replacement, because of its complexity, has been considered at length by Yates (1960) and Patterson (1950).

Sampling on Two Occasions With Partial Replacement

To estimate values of the population mean on two separate occasions, the simplest procedure is to sample the population separately on each occasion. One must follow whatever methods are appropriate for sampling on the particular occasion, regardless of the method used previously. These estimates are overall estimates as per case (2) above. When the two samples are independent of each other, the overall estimate will contain

nearly all the information available. If the second is a subsample of the first, or if there is partial replacement of the original sample in the second, the matter of utilization of available information becomes more complex.

If the second sample is a subsample of the first, the simplest estimate of a change in $\bar{Y}$ will be obtained from elements common to both samples, that is, elements included in the Boolean intersection of the subsample and the original. An estimate of the population mean on the second occasion is obtained by adding the estimated change to the estimate of the value in the original sample. A more precise estimate for the population mean on the second occasion will be obtained using estimates of regression from the first to the second sample. To accomplish this, one calculates regression estimates of the second sample on the first, using sample estimates from the first sample as supplementary information. A more precise estimate of change is the difference between the value found in the original sample and the regressed value found in the second.

The execution of the latter procedure is outlined as follows. Denote values obtained in the first sample as $\underline{x}$ and those in the second as $\underline{y}$. The values belonging to elements included in both are denoted as x' and y', and those included on the first occasion only as x'. If a fraction, $\underline{f}$, of all the units included in the first sample are included in the second, a random sample produces estimates

$$\bar{x} = f\bar{x}' + (1-f)\bar{x}'' ,$$

$$\bar{y} = \bar{y}' + b(\bar{x}-\bar{x}')$$

$$= \bar{y}' + b(1-f)(\bar{x}''-\bar{x}'),$$

where $\bar{x}$ is the overall estimate for the first occasion and $\bar{y}$ is the regressed estimate for the second. The change from occasion one to occasion two is estimated by

$$\bar{y} - \bar{x} = \bar{y}' - \bar{x}' - (1-b)(1-f)(\bar{x}''-\bar{x}') \ .$$

Calculation of the sample regression coefficient b is based on the values of the elements included on both occasions.

The sampling error associated with $\bar{y}$ then becomes

$$V(\bar{y}) = \frac{V(\bar{y}') - b^2(1-f)\ V(x)}{fn} \ ,$$

where r is the correlation coefficient between values in the first and second samples. The variance associated with the change $(\bar{y}-\bar{x})$ becomes

$$V(\bar{y}-\bar{x}) = \frac{V(\bar{y}') + \left(f-2fb - (1-f)b^2\right)\ V(x)}{fn} \ .$$

If r, the sample correlation coefficient, is close to +1; that is, if the values in the first and second samples are very similar, b will be nearly unity. This is usually the case under a scheme of subsampling. When this is the case, the overall change is nearly that between elements common to both samples, or

$$\bar{y} - \bar{x} \doteq \bar{y}' - \bar{x}' \ .$$

Under a scheme of having both samples the same size, when a fraction, f, is retained and (1-f) is replaced, the previously discussed procedures

may be applied to obtain a sample estimate ($\bar{y}_1$). Another estimate ($\bar{y}_2$) equal to $\bar{y}''$ will be derived from elements included in the second sample only; that is, those used to replace the n(1-f) elements discarded from the first sample. A weighted mean of $\bar{y}_1$ and $\bar{y}_2$ will yield the most accurate estimate $\bar{y}_w$ of the population mean. The weights, w_1 and w_2 will be

$$w_1 = f/(1 - (1-f)^2 r^2)$$

$$w_2 = (1-f)(1-(1-f) r^2)/(1-(1-f)^2 r^2)$$

where r is the correlation coefficient between the values of the elements included in both the first and second samples.

Thus $\bar{y}_w$ is

$$\bar{y}_w = w_1\bar{y}_1 + w_2\bar{y}_2 .$$

where

$$\bar{y}_1 = \bar{y}' + b(\bar{x}-\bar{x}') \text{ and } \bar{y}_2 = \bar{y}''$$

The variance of $\bar{y}_w$, for the case where both samples are of equal size and partial replacement occurs from the first to the second, becomes

$$V(\bar{y}_w) = (1-(1-f) r^2) V(y)/n(1-(1-f)^2 r^2) .$$

with variance

$$V(\bar{y}'_w) = \frac{(1-(1-f) r^2) V(y)}{n' + n'' (1-(1-f)^2 r^2)} .$$

Here n' is the number of elements resampled on the second occasion, n" is the number of new elements, and (1-f) is the proportion of elements sampled on the first occasion but not on the second.

An estimate of the change can be obtained using the weighted average of the change ($\bar{y}'-\bar{x}'$) discussed previously and that discussed above ($\bar{y}"-\bar{x}"$) with weights w' and w" where

$$w' = f/(1-(1-f)r) ,$$

$$w" = (1-f)(1-r)/(1-(1-f)r) .$$

Hence

$$\text{Change} = w'(\bar{y}'-\bar{x}') + w"(\bar{y}"-\bar{x}")$$

with

$$V(\text{Change}) = (1-r)(V(y)+r(x))/n(1-(1-f)r) .$$

This estimate of change is not what one would be lead to believe. The natural choice would seem to be the difference between $\bar{y}_w$ and the overall estimate on the first occasion. The reason is that a more accurate estimate for the population mean on the first occasion is possible once the second is taken. The information on the second occasion as supplementary information. If this adjusted estimate, call it $\bar{x}_w$, is made, the difference $\bar{y}_w - \bar{x}_w$ is approximately the change formulated above. The estimate for change above, has as a condition that the variance on both occasions must equal. If they are not, it is not the more accurate of the two, but equality of variances is usually a reasonable assumption.

When the correlation is perfect between the elements in the first and second occasions; that is, r=1, the formula for change above becomes

$$\bar{y}' - \bar{x}' .$$

which is the difference in values for elements included in both samples. If there is no correlation between the values of the elements on the first and second occasions, the change becomes the difference of the overall means on the two occasions. In this case, also, $\bar{y}_w$ will be the overall mean on the second occasion. And if the values of each element are unchanged from the first to the second occasion, that is ($\bar{x}' = \bar{y}'$), r=1, b=1, than $\bar{y}_w$ is the mean of all values in both samples when each value is included only once.

In regard to the estimation of the regression coefficient b, if the assumption of the equality of variance is reasonable the regression of x and y and that of y and x are both equal to the correlation coefficient r. It is best to replace b by r wherever b appears in the formulae, as r is less sensitive to errors of estimation.

Sampling on h Occasions With Partial Replacement

In the preceeding discussion formulae were developed to accomodate the case when the experimenter sampled on two occasions to estimate the amount of change from one occasion to another. The discussion to follow considers the work of Patterson (1950) who developed procedures to sample on more than two occasions with partial replacement. If no replacement is allowed, the estimates for change discussed previously will apply in the general case. The condition that the two occasions sampled are consecutive can be modified to mean that several appropriate occasions may occur

between the two actually sampled. With the case of partial replacement, which is the most reasonable and often encountered situation, no such simple generalization is possible from the two sample case to the multi-sample case.

Before developing the methods of analysis for sampling on successive occasion, conditions for the efficiency of the estimate must be set forth. Suppose a group of sample elements are divided into sets of possibly unequal size. Denote the i-th element of the j-th set as y_{ij} with set $\underline{j}$ having n_j elements and a total of $\underline{h}$ sets. The population mean in the m-th set, $\bar{Y}_m$ is estimated by $\bar{y}_m$ defined by

$$\bar{y}_m = \sum_{i=1}^{n_j} \sum_{j=1}^{h} w_{ij} y_{ij} ,$$

with the condition that

$$\sum_{i=1}^{n_j} w_{ij} = 1; \quad j = m$$

$$= 0; \quad j \neq m ,$$

which condition minimizes the variance of $\bar{y}_m$. Minimizing

$$V(\sum_i w_{ij} y_{ij}) - 2 \sum_j (k_{mj} \sum_i w_{ij})$$

leads to a set of $\underline{h}$ equations where the undetermined constants k_{mj} are defined by

$$k_{mj} = \text{Cov } (y_{ij}, \bar{y}_m)$$

for all i and j. Three necessary and sufficient conditions for an estimate of $\bar{Y}_m$, E_m, to be efficient are:

(1) Cov $(y_{ij}, E_m) = k_{mj}$ for all i and j,

(2) E_m is an unbiased estimate of $\bar{Y}_m$,

(3) E_m is a linear function of the y_{ij}.

The above conditions will be used to derive estimates when sampling on successive occasions with partial replacement of elements. Assume equal variances on all occasions. Assume that partial correlation coefficients for values of an element more than one occasion removed are zero so that the coefficient of correlation for values of an element one, two, three, etc. occasions apart are r, r^2, r^3, etc. Consider the case where the fraction of elements retained is f, the same on each occasion. On each of the h sampling occasions numbered successively, n elements are included. So, on the (m-1)st occasion (nf) are in the m-th occasion while (un) are replaced, where (u=1-f). Define $\bar{x}'_{m-1}$ as the mean on the(m-1)st occasion for the values of the nf elements common to occasions m-1 and m. Define $\bar{y}'_m$ is the mean of the values on the m-th occasion for the same elements. Similarly $\bar{x}''_{m-1}$ and $\bar{y}''_m$ are the means of values belonging to the (un) elements not common to both samples.

Using conditions previously established, an efficient estimate of $\bar{Y}_h$, the population mean on occasion h is of the form

$$E_h = a_1\bar{y}_{h-1} + a_2\bar{x}''_{n-1} - (a_1+a_2)\bar{x}'_{n-1} + (1-\phi)\bar{y}'_h + \phi\bar{y}''_h ,$$

where a_1, a_2, ϕ are restricted as follows for j not greater than (h-2)

$$a_1 = r(1 - \phi) ,$$

$$a_2 = 0 ,$$

and $\bar{y}_{h-1}$ is a linear function of the y_{ij}. For (j=h-1), the efficiency conditions impose the following restrictions on the constants

$$u(r(1 - \phi) - a_1) = 0 ,$$

$$(1-\phi_h)\left(\frac{S^2(1-r^2)}{fn} + r^2 V(\bar{y}_{h-1})\right) = \phi_h S^2/un ,$$

where the subscript on ϕ is introduced to differentiate between values of ϕ associated with each h. So E_h is identically $\bar{y}_h$. The most efficient estimator of $\bar{Y}_h$ is

$$\bar{y}_h = (1-\phi_h)\left(\bar{y}'_h + (\bar{y}_{h-1} - \bar{x}'_{h-1})r\right) + \phi_h \bar{y}''_h ,$$

where

$$\phi_h = \frac{S^2(1-r^2) + fnr^2 V(\bar{y}_{h-1}) + \frac{fS^2}{u}}{S^2(1-r^2) + fnr^2 V(\bar{y}_{h-1})} .$$

The associated variance is

$$V(\bar{y}_h) = \phi_h S^2/un ,$$

with a similar expression for the variance associated with $\bar{y}_{h-1}$.

The equation for ϕ_h may be rewritten as

$$1 - \phi_h = f/\left(1-(1+f+f\phi_{h-1})r^2\right) .$$

Under the initial condition that ϕ_1 is $\underline{u}$, and given $\underline{r}$ and $\underline{u}$, ϕ_h may be calculated for successive values of $\underline{h}$. Algebraic expressions for ϕ_h may be obtained, the simplest being that for (h=2) which is

$$\phi_2 = u(1-ur^2)/(1-u^2r^2)$$

The limiting value of ϕ_h is reached when $(\phi_h - \phi_{h-1})$ is zero, or $(\phi_h = \phi_{h-1} = \phi)$ at the limit so that ϕ, the value of ϕ_h at the limit becomes

$$\phi = \frac{(r^2-1) + (1-r^2)(1-r^2(1-2f)^2)}{2fr^2} .$$

The limit is reached within the second or third occasion, allowing one to use $(\phi = \phi_2)$ instead of calculating ϕ_h for $\underline{h}$ larger than 2. This shortcut generates two types of error. The weights a_1, a_2, ϕ, used to calculate $\bar{y}_h$ will be slightly incorrect, giving rise to a loss of information. A more serious kind of error is propagated from the fact that the variance of $\bar{y}_h$ is proportional to ϕ.

The fractional loss of information is $(1-\phi_h/P_h)$ and the fractional bias is $(\phi S^2/un)$ the variance (which is too small anyway) is $(1-\phi/P_h)$. This is

the case when the true variance is $(P_h S^2 / un)$ where

$$fP_h = (1 - \phi)^2 \left[u(1-r^2) + fr^2 P_{h-1}\right] + f\phi^2 ,$$

when P_2 is ϕ_2 if the correct weights a_1, a_2, ϕ are used or P_3 is ϕ_3 if such is not the case.

Since estimates of $\bar{Y}_h$ will be calculated at each occasion for sampling, the most logical choice for values to detect change will be $\bar{y}_h$ and $\bar{y}_{h-1}$. The variance of their difference, the change in $\bar{y}_h$, is

$$V(\bar{y}_h - \bar{y}_{h-1}) = 2\phi S^2 (1-(1-\phi)r)/un ,$$

for two consecutive occasions. For two occasions not consecutive, $\bar{y}_h$ and $\bar{y}_{h-k}$, when (h-k) is large enough so that ϕ_{h-k} is ϕ, the variance in the limiting case is

$$V(\bar{y}_h - \bar{y}_{h-k}) = 2\phi S^2 (1-(1-\phi)^k r^k)/un .$$

For future use, ${}_h\bar{y}_{h-k}$, the efficient estimate based on observations from the h occasions could have been used in place of $\bar{y}_h$.

Most Efficient Estimates

The efficient estimate of $\bar{y}_{h-1}$ may be found for the case when there are h occasions sampled. Denote this by ${}_h\bar{y}_{h-1}$, the mean of (h-1) occasions when h occasions have occurred. This serves to avoid confusion with $\bar{y}_{h-1}$ which is the mean of (h-1)-st occasion. Consider a linear function of the variables used to estimate $\bar{y}_h$. Recall

$$\bar{y}_h = (1 - \phi_h)\left(\bar{y}'_h + r(\bar{y}_{h-1} - \bar{x}'_{h-1})\right) + \phi_h \bar{y}''_h$$

with the same criteria on the variable as before; namely, that $\mathrm{cov}(y_{ij}, E_m) = k_{mj}$ and E_m is an unbiased estimate of $\bar{Y}_m$ with E_m a linear function of the y_{ij}. The most simple expression leading to ${}_h\bar{y}_{h-1}$ is

$$E_{h-1} = \bar{y}_{h-1} - w\bar{y}_h + w\bar{y}''_h .$$

Upon application of the criteria, E_{h-1} is ${}_h\bar{y}_{h-1}$ if

$$w = r\phi_{h-1} .$$

The efficient estimate of the mean of (h-1) occasions when h occurred is

$${}_h\bar{y}_{h-1} = \bar{y}_{h-1} - r\phi_{h-1}(\bar{y}_h - \bar{y}''_h),$$

with variance

$$V({}_h\bar{y}_{h-1}) = \phi_{h-1} S^2(1 - \phi_{h-1}(1 - \phi_h)r^2)/un .$$

An efficient estimate of the change from occasion (h-1) to h is the difference between the efficient estimates on occasion h and occasion (h-1) when h occasions have occurred. Thus

$$\text{Change} = \bar{y}_h - {}_h\bar{y}_{h-1} = (1 + r\phi_{h-1})\bar{y}_h - r\phi_{h-1}\bar{y}''_h - \bar{y}_{h-1} .$$

Its variance is

$$V(\text{Change}) = \frac{\phi_h S^2}{un} + \frac{\phi_{h-1} S^2}{un}\left(1 - r^2\phi_{h-1}(1 - \phi_h) - 2r(1 - \phi_h)\right) .$$

As before, when h is more than 2 or 3, ϕ_h becomes constant and

$$V(\text{change}) = \phi S^2 \, (2 - (1-\phi)(2+r\phi)r)/un \ .$$

The preceeding discussion was concerned with the amount of change occurring from one occasion to the next. In the general case when h occasions have occurred, information from preceeding and later occasions can be used for an efficient estimate of ${}_h\bar{y}_{h-k}$. Once more the efficient estimate assumes the form.

$$E_{h-k} = {}_{h-1}\bar{y}_{h-k} - w\bar{y}_h + w\bar{y}_h'' \ .$$

When the criteria for efficient estimates are applied, E_{h-k} satisfies them if

$$w = \phi_{h-k} \, r^k \prod_{i=1}^{k-1} (1 - \phi_{h-i}) \ .$$

Therefore

$${}_h\bar{y}_{h-k} = {}_{h-1}\bar{y}_{h-k} - \phi_{h-k} \, r^k (\bar{y}_h'' - \bar{y}_h) \prod_{i=1}^{k-1} (1 - \phi_{h-i}) \ .$$

This is a lengthy process requiring many calculations. It might be more economical to give up some efficiency in the estimate and gain some ease of calculation and saving of time and money. These less efficient estimates were considered earlier as $\bar{y}_{h-k}$. The efficiency of the unweighted mean ${}_h\bar{y}_{h-k}$ is $\phi/(2u - \phi)$ while that for $\bar{y}_{h-k}$ is $u/(2u - \phi)$. For high correlation the efficiency of $\bar{y}_{h-k}$ is low and tends to 0.5 as the correlation approaches unit.

The derivation of the variance of the efficient estimate is quite complex and will only be stated as a result by Patterson (1950). When (h-k) is

sufficiently large then ϕ_{h-k}, thus ϕ_h, is ϕ, in the limit and the appropriate expression for the variance of ${}_h\bar{y}_{h-k}$ is

$$\phi S^2 \left(1 - \frac{u-\phi}{2u-\phi}\left(1 - (1-\phi)^{2k} r^{2k}\right)\right)/un; \quad \phi = 0$$

or, when k is large enough

$$V({}_h\bar{y}_{h-k}) = \phi S^2/n(2u-\phi) .$$

When the efficient estimate of the change between occasions, k units apart, is $({}_h\bar{y}_{h-1} - {}_h\bar{y}_{h-k-1})$ the appropriate expression for the variance of the estimate is

$$V({}_h\bar{y}_{h-1} - {}_h\bar{y}_{h-k-1}) = 2\phi\ S^2(1 - r(1-\phi)r)/n(2u-\phi)$$

The experimenter may find it desirable to test for no change, or, equality of means from one occasion to the next. Let x_n be the mean of the h-th occasion. When $x_n = \bar{y}_h$ the estimate R of the correlation coefficient becomes $R = r(1-\phi)$, and for $x_n = (u\bar{y}_h + f\bar{y}_h)$, $R = rf$. Then the terms

$$x_1,\ x_2 - Rx_1,\ x_3 - Rx_2, \text{ etc.}$$

are independent. So

$$z_1 = x_1,\ z_2 = \frac{1}{1-R}(x_2 - Rx_1),\ z_3 = \frac{1}{1-R}(x_3 - Rx_2), \text{ etc.},$$

implies that the z_h are independent variables, each an estimate of the mean of the h-th occasion. The variance of z_1 is v and the variance of the others is $v \cdot (1+R)/(1-R)$, where v is the variance of the x's. Weighting the z's, the

best estimate of the mean $\bar{Z}$ becomes

$$\bar{z} = \frac{(1+R)z_1 + (1-R)(z_2+z_3+\ldots)}{n-(h-2)R} ,$$

and has variance

$$V(\bar{z}) = \frac{v(1+R)}{h-(h-2)R} ,$$

where, in both expressions, h is the number of z's used. If the x's are normally distributed then

$$\frac{S(wz^2) - \frac{S^2(wz)}{S(w)}}{v}$$

is distributed as a Chi-square with (h-1) degrees of freedom where $w = (1-R)/(1+R)$.

Efficient Estimates: A General Case

Thus far, the discussion has been concerned with samples where the variance from occasion-to-occasion, sample size, and the proportion replaced were of equal size. The next step in generalizing the equation for change, is to obtain the estimate and its variance when the above conditions of equality do not hold. The equations derived for $\bar{y}_h$ and $_h\bar{y}_{h-k}$ still hold, except a new value ϕ_h' for ϕ_h must be obtained. Let the number of elements sampled on occasion h be n_h of which n_h' are from occasion (h-1) and n_h'' are new elements. Then ϕ_h' is calculated from

$$\phi_h' = 1 - \frac{n_h' n_{h-1}'}{n_h n_{h-1}'' - r^2 n_h''(n_{h-1}'' - \phi_{h-1}' n_h')} ; \; n_{h-1}'' \neq 0 .$$

The appropriate variance for $\bar{y}_h$, when $n''_{h-1} \neq 0$, is

$$V(\bar{y}_h) = \phi'_h \, S^2/n''_h \ .$$

If $n''_h = 0$ the proper expression of the variance is

$$V(\bar{y}_h) = S^2 \left(\frac{1-r^2}{n_h} + \frac{r^2\phi'_{h-1}}{n''_{h-1}}\right) \ ; \ n''_{h-1} \neq 0 \ .$$

If $n_h = n_{h-1}$ for all h, ϕ'_h is identical to ϕ_h. By using the efficient estimate ${}_h\bar{y}_{h-1}$ as

$${}_h\bar{y}_{h-1} = \bar{y}_{h-1} - w\bar{y}_h + w\bar{y}''_h \ ,$$

where

$$w = r\phi'_{h-1} \, n''_h/n''_{h-1} \ ,$$

the variance becomes

$$V({}_h\bar{y}_{h-1}) = \phi'_{h-1} \, S^2(1 - (1 - \phi'_h)wr)/n''_{h-1} \ .$$

A simpler method was proposed by Yates (1960) to adjust sample estimates when unequal numbers of elements were sampled on different occasions. He suggested using a weight on the mean of the elements not common to both samples with

$$\phi' = n''_h\phi/u'n_h \ ,$$

where u' is the average fraction of elements replaced for the h occasions.

An immediate question concerns the method of selecting the number of elements to be included in a sample and the number of these which are new in the sample when going from occasion h to (h+1). Assume that results up to occasion (h-1) are known. Suppose $\bar{y}_h$ is to be estimated such that

$$V(\bar{y}_h) = V(\bar{y}_{h-1}) ,$$

with cost held to a minimum on the h-th occasion. Since

$$\phi_h'/n_h'' = \phi_{h-1}'/n_{h-1}'' = 1/N ,$$

and the cost of including any element in the sample is the same, and the population variances are equal, then $(n_h' + n_h'')$ is minimized when

$$n_h' = N(1-\phi_h')(1-r^2)/(1-(1-\phi_h')r^2)$$

and

$$n_h'' = N\phi_h' .$$

This leads to a constant n_h', for all h, as

$$n_h' = N(1 - \sqrt{1-r^2}\,)\,\sqrt{1-r^2}\,/\,r^2$$

when $h > 1$. Therefore $n_h' = n_h''$ for all $h > 1$.

The sampling scheme is to choose N elements on the first occasion, and retain n_h' elements, as above, from the preceeding sample, choosing an equal number of new elements from the population for the second. On all succeeding occasions use a (retain: replace) ratio of (1:1), where $f = u = \frac{1}{2}$.

The experimenter may be interested in arranging for a given accuracy in $\bar{y}_h$, as well as in setting some boundary on ϕ_h'. The procedure for sampling in this case will be to choose ϕ_h' such that

$$1 < \phi_h' < n_h' / N .$$

On the second occasion retain n_h' elements, choosing n_h'' new ones, and follow in succeeding samples with the proportion retained as

$$f = \frac{\phi_h'(1-r^2(1-\phi_h'))}{1+r^2(1-\phi_h')^2 - 2r^2(1-\phi_h')} .$$

ROTATION SAMPLING

The preceeding discussions were concerned with the case where the sample on an occasion consisted of a certain number of elements carried over from the preceeding sample and a number of new elements. The values of these new elements, which were used in estimating the mean on that occasion, were taken on the specified occassion.

To be more general, Eckler (1955) considers the case where these "new" element's values are also allowed to be taken from the preceeding sample occasion. Denote the value of the j-th element on the i-th occasion as h_{ij}, with values $y_{h-1,j}$ and $y_{h,j}$ allowed to enter the sample on occasion h. Eckler (1957) termed this rotation sampling, in this case two-level rotation sampling. One level rotation sampling is the name applied to the method of Patterson discussed previously.

The sample on occasion h is built by selecting n elements from occasion (h-1) and h. Using the same technique as Patterson and with the same

criteria for the minimum variance, the iterative form for $\bar{y}_h$ under two level rotational sampling, $\bar{y}_{2h}$, is

$$\bar{y}_{2,h} = \bar{y}_h - \phi_h \bar{y}_{h-1} + \phi_h \bar{y}_{2,h-1} ,$$

where ϕ_h is

$$\phi_h = r/(2-r\phi_{h-1}) ,$$

with $\phi_1 = 0$ initially. The notation is slightly different because of the way the same elements are used in both h-1 and h, with only the values, thus the means, changing. Here $\bar{y}_h$ denotes the mean value of the elements in sample h on occasion h and $\bar{y}_{h-1}$ the mean value of the elements in sample h on occasion h-1. The associated variance of $\bar{y}_{2,h}$ is

$$V(\bar{y}_{2,h}) = \frac{S^2(1-\phi_h r)}{n} .$$

The sequence of ϕ_h converges to ϕ where

$$\phi = \left(1 - \sqrt{1-r^2}\right)/r ,$$

so that in the limiting case

$$V(\bar{y}_{2,h}) = S^2 \sqrt{1-r^2} / n .$$

From two level rotational sampling the next step is to three level. Under this scheme n elements are chosen on occasion h. Their values on occasions h, (h-1), and (h-2) are used to build the sample. To find a

minimum variance, unbiased estimator $\bar{y}_{3,h}$ for the mean on occasion h under three level rotational sampling a linear combination of the $\bar{y}'$ for occasions h, (h-1), (h-2) must be found. The coefficients of the $\bar{y}'$ must be such that not all are zero and the number of these unknown coefficients must be equal to the number of covariance conditions as stated by Patterson (1950). The appropriate estimate is

$$\bar{y}_{3,h} = \bar{y}_h - a_h\bar{y}_{h-1} + a_h\bar{y}_{3,h-1} - b_h\bar{y}_{h-2} - (c_h+f_h)\bar{y}'_{h-2} + (b_h+c_h)\bar{y}''_{h-2}$$

$$- d_h\bar{y}'_{h-3} - e_h\bar{y}''_{h-3} + (d_h+e_h)\bar{y}''_{h-3} + f_h\bar{y}_{3,h-2} .$$

Define $\bar{y}_h$ as the mean of the n sample values on occasion h, $\bar{y}'_h$ as the mean on occasion h of the n values sampled on occasion (h-1), $\bar{y}''_h$ as the mean of the n values on occasion h which were sampled on occasion (h-2) and $\bar{y}''_h{}'$ as the mean on the h-th occasion of those sampled on (h-3). No elements are retained in the sample from one occasion to the next; that is, the n elements sampled on occasion h are not the n elements sampled on occasion (h-1), (h-2), etc., and vice versa.

When the condition for $\bar{y}_{3,h}$ to be a minimum variance unbiased estimator is applied, the coefficients become

$$a_h = r/2 ,$$

$$b_h = r^2\left[(3+r)^2 - 2b_{h-2}(1-r^2)\right]/2\left[(9-r^2) - 2b_{h-2}(3+r^2)\right]$$

$$c_h = (b_h(1+r^2)-f_h)/(1-r^2)$$

$$d_h = -(b_h+c_h)r = -e_h ,$$

$$f_h = (r^2+2b_h)/(3-2b_{h-2}) ,$$

for h greater than 3. The b_h are the only truly unknown coefficients, and Eckler found the limit b of the b_h to be

$$4b = (3-r^2) - \sqrt{(1-r^2)(9-r^2)} \quad .$$

The variance of $\bar{y}_{3,h}$ in the above limiting case, regardless of h, is

$$v(\bar{y}_{3,h}) = S^2 \left[(1-r^2)(4-r^2) + r^2 \sqrt{(1-r^2)(9-r^2)}\right] / 4n$$

$$\doteq \frac{S^2}{n}$$

$$\doteq V(\bar{y}) \ ,$$

which holds unless r is nearly unity. Generalization from three to four, and further, levels of rotation sampling is now somewhat obvious. Now the experimenter must ask how long information is still valid. That is, if one were sampling on economic characters, how valid is the estimate of the average wage based on ten or twelve years previous when these figures are compiled every year. Is it worth the computational difficulties to get an estimate very slightly better than one based on the last two or three years figures? Usually not.

The question of how many levels to use, becomes one of choosing among one, two, or three levels. Assume that it costs c to include a sample value on occasion h and $c(1+k)$ to obtain both values y_h and y_{h-1} for a particular element where $0 \leq k \leq 1$. In other words, cost is not linear in multi-level sampling for values from different occasions for a particular element. Likewise, for three level sampling the associated cost for the three values

of an element is c(1+2k). If the total appropriation for the survey is T, then

$$n_1 = T/c \ ,$$

$$n_2 = T/c(1+k),$$

$$n_3 = T/c(1+2k),$$

where the subscript on n, the sample size, is the level of sampling being consider. The retaining proportion is usually near ½ for one level sampling. To find when one level and two level sampling yield equal precision, equate the variances of the two plans and substitute appropriate quantities for n_1 and n_2 getting

$$k = \left(r - \sqrt{1-r^2}\right)^2/r^2 \ .$$

For k greater than the quantity on the right, use one level rotational sampling, for k less than the quantity use two level to obtain greatest precision.

To decide whether to use two or three level sampling, equate the appropriate variances and substitute the values for n_2 and n_3 as before. The constant k will be used as the critical value for determining which level to use. For k smaller, use three level; for k greater, use two level rotational sampling. For four level or higher rotational sampling to be most advantageous, k must be very small with a high correlation.

ACKNOWLEDGMENT

The author wishes to express his appreciation for the suggestions and assistance provided by Dr. A. M. Feyerherm. These aids were most helpful in preparing this report, as well as in preparing the author academically for future research.

REFERENCES

Cochran, W.G.
Sampling Techniques. New York: John Wiley and Sons Inc., 1963.

Eckler, A.R.
Rotation Sampling. Annals of Mathematical Statistics. 26:664-685. 1955.

Jambunathan, M.V.
A Note on the Efficiency of Double Sampling for Stratification. Sankhya. 22:365-367. 1960.

Jessen, R.J.
Statistical Investigation of a sample survey for Obtaining Farm Facts. Iowa State College of Agriculture and Mechanical Arts Research Bulletin. 304:54-59. 1942.

Neyman, J.
Contribution to the Theory of Sampling Human Populations. Journal of the American Statistical Association. 33:101-116. 1938.

Patterson, H.D.
Sampling on Successive Occasions with Partial Replacement of Units. Journal of the Royal Statistical Society, Series B. 12:241-255. 1950.

Robson, D.S.
Application of Multivariate Polykays to the Theory of Unbiased Ratio Type Estimation. Journal of the American Statistical Association 52:511-522.

_____, and King, A.J.
Multiple Sampling of Attributes. Journal of the American Statistical Association. 47:203-215. 1952.

Sukhatme, B.V.
Some Ratio-type Estimates in Two Phase Sampling. Journal of the American Statistical Association. 57:628-632. 1962.

Tukey, J.W.
Keeping Moment-like Sampling Computations Simple. Annals of Mathematical Statistics. 27:37-54. 1956.

Yates, F.
Sampling Methods for Censuses and Surveys. London:Griffin and Co., 1960.

MULTI-PHASE SAMPLING IN CENSUSES AND SURVEYS

by

CHARLES ALBERT BENDER

B.S., Kansas State University 1965

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Multi-phase sampling is a technique employed to obtain estimates of parameters using information from previous samples. The widest application of multiphase sampling is in censuses and surveys such as those conducted by United States Government agencies. Two problems best resolved using multi-phase sampling are the estimation of the mean value of a character when sampling is difficult or expensive and the estimation of a mean and its change over a time interval.

In the estimation of the mean when sampling is expensive or difficult, double, or two-phase sampling is employed. The technique of double sampling was first set out by Neyman in 1938 using the sample mean as an estimate of the population mean. For this method to produce a more precise estimate than one produced by simple random sampling, the character of interest must by correlated to some other more easily sampled character. The amount of advantage gained by double sampling is in direct proportion to the strength of the correlation and the degree to which the correlated character is more easily or cheaply sampled. Neyman developed methods for allocation of expenditures used to sample on the primary character and on the correlated character which produced minimum variance for a fixed cost. Neyman's allocation method was generalized by Jambunathan in 1960.

In 1962, Sukhatme developed ratio estimates to estimate the population mean. In 1963, Cochran and Cox turned to estimates of the regression of the primary character on the secondary to more fully utilize available information. The gain in precision over simple random sampling using regression estimates is directly proportional to the square of the regression and the degree to which the secondary character is more inexpensively measured.

Double sampling was extended to triple, or three-phase sampling by Robson and King in 1952. Instead of one correlated character, two additional characters are sampled which are correlated to the primary character. Between the two secondary characters there exists a correlation also. As before, there is a hierarachy of cost involved in sampling on the primary character, the secondary, and tertiary, the last two of which are the two additional characters correlated to the first. An application of this is found in the Curtis Impact Survey.

The second general area of application of multi-phase sampling is that of finding a mean and its change over successive sampling occasions. When a population is sampled on several occasions there are three quantities for which estimates are desired: (1) changes in the mean from one occasion to another, (2) the average of the means over several occasions and (3) the mean on the last occasion. Due to its complexity, partial replacement of sample elements when finding estimates for the third quantity above was considered at length. Yates, in 1960, and Patterson, in 1950, derived minimum variance estimates for the mean on the most recent sampling occasion under a scheme of partial replacement of sample elements. They derived several estimates for the change in the mean using information from one or more previous occasions.

In 1955, Eckler considered the case where elements were carried over from the previous occasion to contribute to the sample to the next, but some not sampled on one occasion are included in the next. Double sampling discussed previously became one level rotation sampling and further extensions were made to two and three level.